Study for the Storage of Electronic State Publications

Bullet Process, LLC

Prepared for:

Fingold Library – State Library for the Commonwealth of Massachusetts

June 26, 2003

Bullet Process, LLC ▪ 1169 Commonwealth Ave, Suite 3 ▪ Boston, MA 02134

Table Of Contents

Introduction

Executive Summary

The findings of this study are the following:

- The State Library must establish a central storage repository for electronic agency documents to comply with State Law.
- This project introduces an automated process to capture state documents, which is less labor intensive than manually retrieving and classifying publications.
- The project provides a software platform that can aid other areas of the library with document categorization.
- Software automation is more accurate and less costly than an extensive communications campaign to track all State Agency documents.
- The project allows for compliance with future government indexing initiatives for electronic documents.

Purpose

This document is intended to provide a high-level understanding of the Fingold Library State Library for the Commonwealth of Massachusetts project regarding the Study for the Storage of Electronic State Publications based on the RFR SLM-015. This document evaluates the feasibility of the project and examines the business requirements.

Scope

This study assesses the storage requirements for State Agencies and long-term feasibility and cost of a business solution of State Library storage. Its primary purpose is to consider the project from a one year time horizon sufficient to complete the initial implementation of the project. This project is the first of a multi-phase initiative to automate key library operations

The scope of the study is to develop strategies for collection and long-term access and storage of all types of electronic publications created by Commonwealth's agencies. The State Library has hired Bullet Process to interview the selected agencies about current storage strategies for electronic publications and to make recommendations for the creation of a centralized depository overseen by the State Library.

Methodology

Bullet Process LLC conducted the study for the State Library by conducting interviews with State agencies and through contact with staff at the State Library.

The Bullet Process Approach focuses on understanding business and technical needs by clearly defining key objectives, which simplify the project. Often we find that this decreases client implementation times while exceeding customer service expectations.

Definitions

The following legal definitions for agencies and documents were provided by the State Library:

"Agency": each agency, office, state-supported institution of higher education, department, authority, executive office, bureau, officer, board, committee, task force, commission, special commission, division, executive office of the Commonwealth, whether permanent or temporary in nature, including any state agency supported wholly or in part by public funds.

"Publication," any document, study, rule, regulation, report, directory, pamphlet, brochure, periodical, newsletter, bibliography, micro-photographic form, tape or disc recording, annual, biennial or special report, statistical compendium, or other printed material regardless of its format or manner of duplication, issued in the name of or at the request of any agency of the commonwealth or produced and issued as part of a contract entered into by any agency of the Commonwealth regardless of the source of funding, provided they constitute "public records" as defined in clause Twenty-sixth of section seven of chapter four, excepting correspondence, blank forms, and university press publications.

Business Objectives

Background

According to Massachusetts General Laws (MGL) chapter 6 Section 39 B the State Library is a depository for every state document. Section 39 applies to every document regardless of its format or manner of duplication. Agencies are required to send the State Library 8 copies of every publication created by that agency.

Key Trends

In 1999, the State Library received approximately 3,000 documents from state agencies in printed paper form. Last year the Library received less than half the number received in 1999, roughly 1200. Cost cutting measures in the State have encouraged many state agencies to no longer print their documents.

Most agencies are now transitioning to electronic only publications. However, not all agencies are aware the compliance obligates them to continue to send the electronic documents to the State Library. When an agency printed a document, the state printer would often fulfill the law obligation by sending the State Library 8 copies of the document. With electronic documents, the burden of compliance is on the agency to send their documents.

The State Library aims to address this change and develop guidelines that ensure the collection, storage and permanent access to these publications.

The following business factors are motivating factors of the Library project.

1) Make a lasting archive of state documents available to researchers
    a. The State Library promotes a strong and comprehensive collection of information materials to preserve a historical sense of the Commonwealth. The State Library houses historical records dating back to the 1800s.
2) Compliance with State Law
    a. The State Library has a statutory mandate to preserve state documents.
3) Preserve documents when agencies remove or update the electronic version.

Survey Results

Bullet Process conducted a survey of Massachusetts' agencies to determine the storage needs and issues in creating a central facility for storage of state documents. This section contains the results of face-to-face interviews and a questionnaire from the following state agencies:

1) Operations Services Division (OSD)
    a. OSD facilitates and audits the acquisition of commodities, professional, general, human and social services. These acquisitions support the socio-economic goals of the Commonwealth including disadvantaged business, environmental, and other programs that are in the best interest of the Commonwealth. Additionally, OSD has the responsibility for the administration of printing, fleet operations, and surplus personal property dispositions.
2) Appellate Tax Board (ATB)
    a. The ATB is a quasi-judicial agency within the executive branch "but not subject to its control in the conduct of its adjudicatory functions." MG.L. c. 58A,1. It is devoted exclusively to hearing and deciding cases on appeal from any state or local taxing authority. The board was established by the legislature in 1929 to relieve the Superior Court of the large volume of tax appeals, and to provide taxpayers with a less expensive and more expedient means of appeal. Since the board handles appeals related to virtually all state tax and excises as well as appeals of local property taxes from all 351 cities and towns in the Commonwealth, it is at the center of the entire system of revenue production for the Commonwealth.
3) Metropolitan District Commission (MDC)
    a. The MDC manages the parks, water quality, and certain roadways around Boston and surrounding towns.
4) State Library of Massachusetts
    a. The State Library has maintained a position of vital importance in the State House since its formal establishment in 1826. Starting from a collection of maps, statute books, and government documents, the Library has grown into a multifaceted resource for legislators, executive personnel, state employees, historians, genealogists, and interested citizens. Through political turmoil of one hundred and sixty years in the State House, the State Library has continuously reflected the progressive ideals of the legislative and executive branches of government through its collection of state publications.

Long Term Storage and Access Policy

Each Agency was asked about its policies for the long-term storage of documents. The survey evaluated their policies in creation, storage, and distribution of documents.

Storage of Electronic Documents for Massachusetts Agencies		
Agency	**Current Storage**	**Access Method**
Operations Service Division	Physical, Electronic	Compass, Agency Web Site, Internal
Appellate Tax Board	Physical, Electronic	Agency Web Site, External Legal Publications
State **Library** of Massachusetts	Physical, Electronic, **Microforms**	Agency Web Site, Internal Library Stacks
Metropolitan District Commission	Physical, Electronic	Agency Web Site, Internal Archives

The current storage and access to agency documents is a broad mix based on their operational needs. Each agency has a mix of physical documents that are published, archived, or stored at the agency. Each agency also stores their documents electronically for printed publication or electronic distribution.

Access to the agency documents is also broad and fits in to the agencies operational needs and mandates:

- OSD had the most comprehensive electronic access through Compass and their website.
- Other agencies may rely on 3rd parties, which publish information related to the agency. A 3rd party printer publishes findings of fact for the ATB.
- Not all agency documents are available electronically; the MDC maintains a large history of physical paper-based pictures and records.

Agency Publications

Each agency provided examples of documents and the frequency these documents were updated:

OSD

Document	Update Frequency	Number and Length
Comm-Pass RFRs and related documents	Rarely	All Documents retained
Uniform Financial Report and related documents		Filings from 2002 on
D1 Performance Measures		Filings from 2002 on
Handbook	Update twice per year	1
Policies and Procedures		Periodic Updates
OSD Discussions		
OSD Contract Documents		
OSD Updates	Rarely	60 per year (50k to 1 MB)

Many OSD documents are contract or publications, which do not require updates or revisions. Some are published frequently through the year and others are revised only a couple of times per year.

The overall number of documents at the agency could be estimated at 2 million or more. These documents are a variety of documents available through the web and also staff member documents residing on their office desktops.

ATB

The ATB publishes several agency guidelines per year. They also have email distribution of a newsletter. Lexis/Nexis publishes findings of fact with the full opinion of appellate cases, which vary in size from 10-60 pages, 39 were published for 2002. (Steve's not is "on what?" Give more description about the types of publications are published.)

Document	Update Frequency	Number and Length
Newsletter	No updates	10-15
Forms and Instructions	1-2 per year	

MDC

The MDC publishes:

Document	Update Frequency	Number and Length
Master Plans	Every 25-50 years	10-15
Watershed Access Plans	Every 5 to 10 years	< 50
Watershed Circulars	As needed	
Facility Guide	Every few years	

The MDC maintains documents that are updated infrequently and are generally reference oriented.

Distribution Strategies

Distribution Strategies for Electronic Documents			
Agency	**Electronic Distribution Method**	**Web Access**	**Email List Distribution**
Operations Service Division	Web Publishing, Email with or without attachments	Yes	Yes
Appellate Tax Board	Web Publishing, Email with or without attachments	Yes	Yes
State Library of Massachusetts	Web Publishing, Email with or without attachments	Yes	Yes
Metropolitan District Commission	Web Publishing, Email with or without attachments	Yes	Yes

All of the agencies use the web and email for distribution of their documents. This does not restrict any of the internal storage needs of the agencies. Each agency stores the original electronic document for internal operations for the statutory life of a document. In addition, the agency makes a copy of the document available on their website or via email distribution.

In discussing the agencies storage needs, there was no indication that their websites had too many documents. The website was only one part of their distribution strategy. Documents that frequently changed are often emailed which has very little or no permanent storage requirement for the agency. If the agency did have a website that was less content dynamic, as was the case with OSD's compass, it relied on a database for much of the changing content.

A finding of the study is that a major effort that the State Library needs to address is documents sent via email. The agencies use email distribution lists as a tool for mass distribution of documents. Each agency had a different purpose with their email lists and in some cases there were sensitivity issues around the recipients of the email.

Not all agencies were aware that they could simply add the State Library to their email list to gain compliance with State law. Agencies can add an email address from the State Library email address to their email lists and the Library is sent a copy of the publications. This is one method of collecting publications the State Library presently uses to collect new publications and be notified about other materials in the agency's website.

Document Authoring

Document Authoring for Electronic Documents				
Agency	**Creation Mechanisms**	**Realtime Web Publishing**	**Document Formats**	**Hardware**
Operations Service Division	MS Office, HTML, Web Publishing Package (Teamsite)	Yes	Microsoft Office, HTML, PDF	Internal Facilities, PCs, Servers
Appellate Tax Board	MS Office, HTML	Limited	Microsoft Office, HTML, PDF	Internal Facilities, PCs, Servers
State Library of Massachusetts	Yes	Limited	Microsoft Office, HTML, PDF	Internal Facilities, PCs, Servers
Metropolitan District Commission	Yes	Limited	Microsoft Office, HTML, PDF	Internal Facilities, PCs, Servers

Each agency uses different software to create their electronic documents but all use a consistent set of standard file formats. The output formats were html, pdf, and Microsoft Office related (i.e. .doc or .xls). PDF and Microsoft Office related files are created using their respective programs. HTML was done in a variety of ways from professional content authoring packages to a simple text editor.

A finding of the study is that there is not a unified document management methodology used by all Commonwealth agencies. Every agency has developed its own individual procedures for dealing with the preservation of materials. When asked about how the agencies create their documents there was very little workflow required to create and publish documents. This was reflected in how documents were stored. Most agencies relied on the staff member storing their documents on space on their office desktop or on a network file server.

The State Library will not be able to use an integrated software solution, which would link all the document management solutions at each agency. There simply is not this kind of infrastructure consistently available at every agency. Instead, the State Library will need to rely on the agency communicating when documents are changed or updated. The State Library should evaluate software solutions which help manage this communication by transparently examining the website which reflects the official (latest) version of an agency document.

Archival Strategy

Archival Strategy for Electronic Documents		
Agency	Archival Requirement for Electronic Documents	Archive of Staff Documents
Operations Service Division	Generally 7 years	Staff Discretion
Appellate Tax Board	Staff Discretion	Staff Discretion
State Library of Massachusetts	Staff Discretion	Staff Discretion
Metropolitan District Commission	Staff Discretion	Staff Discretion

Much of the archival strategy of the agencies is left to the staff member's discretion. One staff member at OSD has a very specific method of archiving her documents from year to year to keep with compliance issues. But not all agencies maintain, are mandated, or sponsor a policy such as this. Much of the archival strategy is dependent on individual working styles, computer equipment, and backup procedures at the agencies. Lacking a standard policy for archival and backup of documents agencies perform these tasks at their own discretion.

Agency Staffing

Agency Staffing for Electronic Documents for Massachusetts Agencies				
Agency	Dedicated Staff sending library documents	Agency aware of compliance regulations?	Is Agency advising library of changes of electronic documents?	Agency Staff Full-time for document archival?
Operations Service Division	Yes	Yes	No	No
Appellate Tax Board	Yes	Yes	No	No
State Library of Massachusetts	Yes	Yes	No	No
Metropolitan District Commission	Yes	Yes	No	No

The agencies communication with the Library is based on the established and common practice for print documents. They were all aware they are required to send 8 copies of a printed document to the library, which was often fulfilled by the State printer. They were unaware that their electronic documents were subject to the same policy.

The staffing at the agencies reflected this communication objective. Each agency had a specific person, which communicated with the library that new print documents were published or available. It's possible that the person responsible for print documents may also be responsible for electronic documents but this is not always the case. Some agencies have a full-time staff member dedicated to print publications and a separate webmaster who is responsible for online publishing.

In all cases, there was not a full-time person responsible for communicating new documents or revisions to the State Library. The cost to each agency would make this prohibitive especially for smaller agencies. The State Library should look towards a software package which can transparently achieve what a full-time staff member could at a fraction of the price to the Commonwealth.

Survey Trends and Common Problems

In talking with the agencies there were a number of trends and common problems, these are summarized below:

- Communicating to the Library when a document is removed
- Communicating to the Library when a new electronic document is added
- Communicating to the Library when a new revision of a document is made available

The State Library requires a transparent mechanism to collect the documents from agencies. The State Library does have a relationship with the State Printer which makes it easier to collect some of the publications produced by the agencies. Developing a system that automatically sends an electronic copy to the State Library would mean that more agency information can be preserved. The State Library needs an electronic equivalent.

- Dead Links on website

Agencies with big websites at one time or another have dead links (ie web links that do not point to a valid document). State initiatives such as the portal have reduced these but not entirely eliminated the problem. In addition, when an agency removes a document any links the State Library has already captured would become dead links. This underscores the importance of maintaining a central repository of documents to preserve them.

- Agency captures information to a database which is not readily accessible for archive

Not all agencies have all of their internal content in a readily accessible format to capture.

- Agencies have different needs in internal archiving of their documents. Not all documents are permanently archived as they are kept on personal machines (i.e. not servers)

Each agency had a different statutory requirement for archive and policies for storage of documents on office desktops and network file servers.

- Some documents require a signature

Some electronic documents are stored but not complete as their paper counterparts. Often contract documents are stored in an electronic form and archived on-line but they lack a signature. The live version of a contract might be stored in another way. For historical archiving purposes, the paper document contains some historical value.

- Web documents are treated differently then print documents

Electronic documents on the web require a more automated process to store and maintain than print documents. Print documents are easy to capture as each revision is send to print. With an electronic document there are very limited obstacles to posting a new version on the website. As a result, the State Library can take a communication-oriented approach where they are notified of every change or use software which automates this process.

- Agencies post web documents themselves, whereas they rely on the state printer for printed documents. Agencies use different creation and publishing software for electronic documents

There is no uniform methodology for document publishing for every agency. Some agencies rely on external IT resources, while some have dedicated internal staff. Each agency selects creation and publishing software for use with its documents. These software packages vary widely between agencies.

- Agencies are not printing as many documents as cost savings measures rely more on electronic distribution
- Agencies do not generally mandate a policy for long-term archival of personal or non-distribution documents

Return On Investment (ROI)

The Return On Investment for the project is believed to be a significant return based on the following benefits to the Commonwealth of Massachusetts:

- Ease the burden of Compliance with State regulations
- Introduce an automated process to capture state documents, which is less labor intensive.
- Provide a software platform that can aid other areas of the library.
- Software automation is more accurate and less costly than an extensive communications campaign to track all State Agency documents.
- Allow for compliance with future government cataloging initiatives for electronic documents.

Ease the burden of Compliance with State Law.

The software platform will serve as a central platform in easing the agencies compliance burden and building a long-term archive of documents.

- Each agency does not have full-time staff to dedicate to creation of the State Library document repository.
- The project will allow the agency to keep pace with the changes and support electronic document publishing.

Introduce an automated process to capture state documents, which is less labor intensive.

The number of documents required to be captured and stored could be very labor intensive and expensive for the State Library to undertake over a long period of time. Electronic documents may require manual classification. The project would serve to ensure the State Library could operate a large repository. The project can also be adapted to future needs.

Provide a software platform that can aid other areas of the library.

The project is scalable for future needs that require advanced indexing and storage of electronic documents. It could be used for other state purposes besides storage of agency documents.

Software automation is more accurate and less costly than an extensive communications campaign to track all State Agency documents.

The State Library currently runs an extensive grassroots campaign to learn about and acquire agency documents. Automated software can ensure that a much

higher number of the agencies could be reached and supported. The software can more transparently capture agency documents requiring less effort from each agency.

Allow for compliance with future government cataloging initiatives for electronic documents.

The State Library will participate in future cataloging initiatives as state or federal laws may mandate certain metadata standard be implemented. This project will create a basic method for capturing metadata information, which can be applied towards any cataloging standard. This will result in cost savings over manual reclassification of data to suit new purposes.

Process Objectives

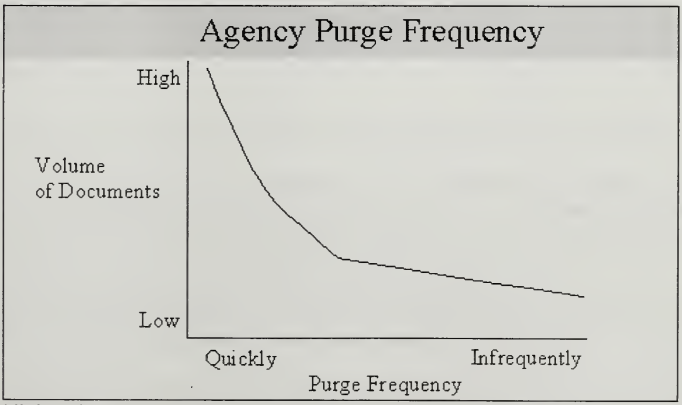
The study examined external and internal factors, which affect the State Library process for collecting documents. This section explores those factors, evaluates the library problems for this project, and examines the software required to achieve the process objectives. The process is also affected by the critical risks of the project, which are addressed in the project implementation section.

Agency Factors

- Agencies long term storage requirements are closely dictated by statutory concerns.

The State Library process is affected by the length of time an agency is required to keep a document. Statutory concerns are usually addressed by the agency by keeping the document available for that period. Once this requirement is satisfied the agency will usually archive the document and remove it from an operational context. This requires the State Library process to be flexible in the number of documents it accepts from an agency. As statutes change, so will the storage requirements. This might lead to a quick influx of documents.

- Agencies purge their documents at different rates.



High volume of documents are purged at regular intervals (i.e. 7 years)
Medium volume of documents (long term) none found
Low volume (long term), website or internal

This chart illustrates the general pattern that agencies with high volume of documents tend to purge them more frequently than agencies with lower volume of documents. The State Library process should account for agencies with documents that change or add new documents frequently. The process would be best designed to allow for automation rather than Library staff trying to receive communication on every document change for every document an agency publishes.

Library Factors

- Resolve agencies required compliance obligations in furnishing documents to library

The State Library process should be a reliable way for agencies to send their documents and ensure that they are in compliance with state law. The State Library currently produces a quarterly checklist of printed publications. If an agency wished to verify that the library received their documents they could consult this list. The electronic process should also be transparent in this way.

- Provide a widely used mechanism that agencies can easily use

The State Library process should support a transparent mechanism to receive documents from agencies without overburdening communication with the agency. The process should support agencies contacting the State Library with documents or revisions without requiring additional staff time or manpower. A non-automated process could result in thousands of hours of additional work for the State Library as the volume of electronic documents grows.

- Web based submission and document changes can be automated.

The mechanism should be transparent as the state printer is to the agencies now. The State Library process can also support the agencies in new ways. The new process can allow the State Librarians to scan the agencies website for documents and to automatically retrieve updates. A paper-based approach would be cost prohibitive to query the agencies in this manner. With the electronic process it is possible and desirable. It allows the library to maintain a more complete archive then previously possible.

- Simple mechanism for agencies to add new document
- Provide multiple ways documents can be accepted
    - Agency posting to web site
    - Email to state library

The State Library process should be simple for an agency to add a new document. The process should allow all the normal channels of distributing document to work with the State Library. Email documents sent to the State Library should automatically be added. Web documents posted to the website should automatically be added. Email lists that the agencies have should simply require the State Library email address be added to them.

The mechanism such as emailing to a State Library address can easily be understood and implemented by all agencies. Expensive software document management methodologies are not being used at agencies. If they were an automated integration might be considered. But would be significantly more complex and beyond the scope of this project.

- Provide ways for agencies to specify inclusion or exclusion of documents

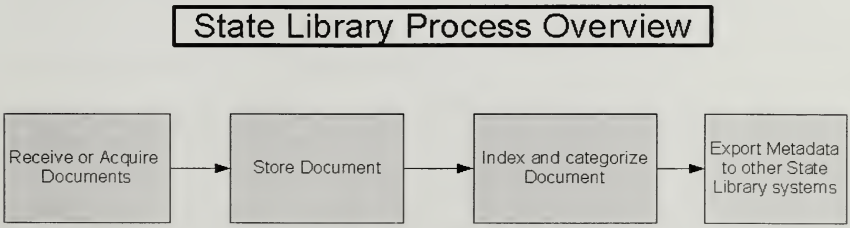
As the State Library process gains acceptance, agencies will need to decide what types of publications should be sent to the State Library. For example, web pages that cannot be indexed for statutory reasons or pages that are dead links. The process should support removing a document from the repository and from indexing.

- Allow for future compliance with state cataloging initiatives

The process and software selected should support future government cataloging initiatives. This may include future state meta-data tagging goals or taxonomy standards.

Proposed Process

Based on the process objectives, it is recommended that the State Library utilize the following process:



1) Receive or acquire documents from a variety of sources

- Email or Email List
- Web Sites
- Internal File Systems

A document is received from agency via email which can include being a recipient of an email list, an agency web site, or an internal State Library file system. The software will need to support scanning an agencies web site for document.

Later phases of the project may also incorporate databases sources to assemble a document which will be archived.

2) Store the document

- Repository of Central Documents
  - URL linking available, as it's a permanent repository of documents.

The State Library will add the document to the central storage repository. In saving the document, it will also backup the file to an electronic medium for permanent storage. The central repository will store a number of years of documents, which will be determined by the library and affected by the cost of the online storage.

3) Index and categorize the document

The State Library will index documents it receives allowing for public search of the metadata created or indexed. The public search and other library systems will utilize the categorization and also provide full-text search of the documents.

This step is the primary differentiator between software packages. All software selected for review in this study have extensive facilities, which will reduce operating costs for the library.

4) Export Metadata to other library systems

The process will allow the metadata created as part of the indexing stage to be exported to other library systems.

Alternative Approaches

This study examined different approaches towards meeting the goals and requirement of this project. The two considered approaches were whether to:

1. Evaluate solutions based on a keyword search methodology and the ongoing maintenance costs.
2. Evaluate solutions based on a natural language methodology and the ongoing maintenance costs.

The second approach was adopted in this study as the most cost effective way for the agency to meet the ROI objectives. The keyword search methodology was determined to have more upfront costs in implementing a system and higher operational costs.

Software Comparison

The software the State Library selects will become an integral part of the process for submitting and storage of electronic documents. The leading software packages were examined in this study based on the following software requirements.

Software Requirements

- Rules Based Organization

The software package must support rules based categorization, which can be defined and changed without requiring extensive manual classification of each document.

- Ability to index unstructured files in a variety of file formats

The software package must support unstructured files in formats including Microsoft Office (Word, Excel), Adobe PDF, and HTML. The agencies have a broad variety of format that contains minimal initial metadata.

- Non-programming related solution for library staff

The software package must support the library staff in making changes to the classification of documents without extensive programming knowledge. The software will support most changes through a software tool. For meta-data that requires manipulation the programming should be ease for staff to learn.

- Ability to serve as a long-term library software platform for use in other library projects

The software package will support a variety of import/export mechanisms that allow for the solution to be integrated with other library systems and to serve as a centerpiece for electronic documents. The software should support file format standards such as XML, CSV, text files, and HTML. The software should also have an API that allows it to be programmatically used with other library software. The API should have a JAVA module.

- Web crawl of agency websites

The software package must support scanning an agencies website for documents. The software will support specifying the frequency of scans and include/exclude documents. Implicit in this approach is an ability to address document revisions. The central repository will capture all revisions of a document and latest one will be indexed and available for full text search.

- Ability to import documents from email.

The software package must support email and email attachments to allow for documents to be added from email. The software should have a module available for Microsoft Exchange or have completed customer implementations using this mail platform. Microsoft Exchange is the mail platform for the State Library.

Software Packages Considered

Blue Angel
(http://www.blueangeltech.com)

The Blue Angel package includes modules for web crawling, categorization, web-based search and metadata management. It includes functionality to assist with workflow management.

It has a best of bread approach relying on Oracle for its full-text document repository and external tools for data transformation.

Search Capabilities: Field Search (including Booleans), stemming, and numeric search.

Categorization is based on the metadata provided, not from unstructured content in the document. Categorization can use related terms and synonyms.

The package will require software extensions to support email.

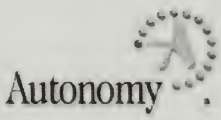
Endeca
(http://www.endeca.com)

The Endeca package includes modules for web data browsing and search, categorization, and metadata creation rules. The Endeca tools enable creation of rules for categorization and navigation. The web based user interface without any customization allows many advanced features to be utilized. With Endeca Guided Navigation, a user will not see categories that contain no documents.

Search Capabilities: Full Text including stemming, guided search, numeric search, phonetic spelling correction, thesaurus, relevance, and suggested alternatives.

Categorization is based on metadata provided and can provide for categorization based on the document. The package can generate taxonomies in a highly automated fashion.

The package has customers that import Exchange based email into the suite but does not have an Exchange connector module.



Autonomy
(http://www.autonomy.com)

Autonomy is a software technology which allows for search, categorization and metadata creation. It does include a limited user interface which is usually customized as part of a solution.

Search Capabilities: Full-text search, parametric search, stemming, Thesaurus, relevance.

Categorization is based on metadata provided but also the unstructured content in the documents. The rules-based engine can pick conceptual terms from the document and generate a taxonomy based.

The package has an Exchange connector module.

Software Recommendations

During the implementation of the selected software product, the State Library should address key software decisions. This section addresses these at a high level.

- Decide on an indexing method to be employed for stated documents.

1) Library Meta tagging: The Library would create meta-tag information for each document or agency acquired or received. This would not involve any agency resources or input.
2) Agency Meta tagging. The State Library could provide, assist, or implement a meta tagging standard with each agency to help categorize their documents.

Once the document has been retrieved there are alternate approaches to categorizing them in to one of the library standards

1) Learned approach: Documents are compared against one of the libraries standards (i.e. GILS or Dublin Core). Documents are then automatically assigned to the taxonomy. The State Library Cataloger can review the categorization choices and further train the system
2) Iterated approach: The State Library Cataloger would create an initial taxonomy from the documents (automated). Which the software, refines the process further using a tool to create taxonomies. These taxonomies would then be normalized against any library standard.

Product Overview

This section addresses the assumptions, costs, and license issues with the project.

Assumptions and Dependencies

The assumptions for this suite of products are somewhat limited as the agency will be the initial user of the suite and also obtain funding for it. There is very limited reliance and dependence on external requirements.

- The initial project will allow Internal and Public access at the State Library's discretion.
- The project will assume that between 5 to 10 concurrent users
- The State Library project consists of two major components a software package platform and a central document repository.

Cost and Pricing

The cost for the project is based on the State Library funding initial one-time costs, ongoing maintenance costs, and operational costs.

The initial funding for the suite includes hardware/software required for a full deployment of the project.

- The agency may need to incur some one time costs for the research and development of the project. These costs will include software and hardware required for testing of the suite or one-time purchases of software development licenses.
- The agency may choose to provide a quick start program arranged through a state vendor to ensure the project can meet aggressive deadlines. A quick start program provides the basic hardware, software, and access to begin development work on the project.
- The agency will require customization of the software selected to meet the specific needs of the State Library.
- The agency may choose to pay for some of the ongoing maintenance costs of the software selected, which may result in a lower overall cost.

The initial funding includes consulting resources required for implementation of the project.

- The consulting resource should act as a primary point of contact for the business and technical aspects of the project. The consulting resource must understand the business requirements of the agency. Ideally the

relationship would add critical technical knowledge on ways to simplify or reduce the cost of the technology implementation.
- The consulting resource will produce key documents needed to fully understand the business requirement

The initial funding for this project includes training of the agency staff with the product in a train- the- trainer model. The consulting resource will work with the agency to develop a training curriculum. The consulting resource would then train an initial group to operate the software.

Licensing and Installation

The software products will require the State Library to purchase the software through a value- added reseller of the software vendor.

Project Implementation

This section of the document addresses the costs and risks associated with the project.

Projected Costs

The projected costs for this project are presented are for several commercially available software packages also including consulting, required hardware, and hosting in the Chelsea data center. The costs are based upon the staff generated requirements, their difficulty, and market rate for services.

Total Project Costs Over First 5 Years							
	Initial	Year 1	Year 2	Year 3	Year 4	Year 5	Total
Blue Angel	$ 331,335	$ 121,577	$ 122,489	$ 124,314	$ 127,964	$ 131,614	$ 959,293
Endeca	$ 232,835	$ 75,949	$ 76,861	$ 78,686	$ 82,336	$ 85,986	$ 632,653
Autonomy	$ 387,835	$ 62,949	$ 63,861	$ 65,686	$ 69,336	$ 72,986	$ 722,653
Data Volume (GB)	10	50	100	200	400	600	

The annual operational costs of this system range from $62,949 to $131,614 based on the number/volume of documents and all related costs (i.e. staffing, software maintenance, and hosting costs). This proposal is based on staff requirements generated for the study.

Blue Angel requires a database package, which will be hosted at the Chelsea data center. The estimated cost for the Oracle license was $80,000 with ongoing costs of $35,000 per year.

The State Library must be able to support this level of ongoing funding commitment to the project or reconsider. The State Library should consider these options going forward as the ideal project would have higher one-time costs with a much lower ongoing cost.

- The data center operational costs can vary between $10,000 to over $50,000 per year.
- The data center operation costs will rise with the level of documents, which could substantially escalate beyond $10,000 per year.

Project Phases

Budget

The projected budget for this project is presented as a whole including all consulting, training, hardware, software, conversion, and data entry.

Phases

Phase	Goal	Description
Phase A	Purchase all hardware and software.	Phase A will provide exact specifications for data center hardware purchase and acquire all required testing hardware. The State Library will also finalize a software purchase.
Phase B	Requirements Analysis	Phase B will further document the full software extensions and process required for a successful software implementation. Library staff will be consulted to ensure other required metadata and library standards are addressed.
Phase C	Software Installation and Taxonomy generation	Phase C will install/configure the software for State Library use, configure the initial metadata, and create an initial taxonomy.
Phase D	Software Extensions	Phase D will implement the required software extensions to support the library process.
Phase E	Integration with other Library Software systems.	Phase E will integrate the project software with other library software packages.

Budget beyond Fiscal Year 2004

Recurring Costs

- Software
  - Each software product requires maintenance costs between $15,000 and $55,000.
  - If Blue Angel is selected it requires a database package, which will be hosted at the Chelsea data center. The estimated ongoing cost for the Oracle license is $35,000 per year.
- Data Center
  - The data center operational costs can vary between $10,000 to over $50,000 per year.

  - The data center operation costs will rise with the level of documents, which could substantially escalate beyond $10,000 per year.
- Hardware
  - Server replacement in 5 years
  - Service contract for all hardware units to ensure for replacement value
- Software support
  - Technical support contract at least 5x40 hours to support all application related issues.

Risks

Functional Risks

- Difficulty in being a domain expert in each agency's documents

The process should not require the staff to be domain experts for every agency. The process should support a mechanism to acquire new agency documents and refine the results of this process based on defined rules. This prevents the Library from having to learn the website of each agency, the mission/objective of each document, and the content matter.

- High barrier to storing paper or electronic documents for 50+ year timeframe

Long term archiving of documents paper or electronic, can be difficult and expensive. Currently the State Library employs only a film and paper based approach. The State Library will need to develop a strategy for electronic storage, which may be applied to other non-electronic documents. For example, the Library may consider imaging some of the film-based documents and also storing those on an electronic medium.

Business Risks

- Electronic documents could result in additional staff commitment and manpower beyond acceptable levels.

The State Library should select a software solution which allows for a very highly automated process. It will be very expensive for State Library staff members to repeatedly handle a single electronic document. The process should support a document being submitted and the only required staff interaction is actually cataloging the document. A potential problem is requiring an electronic document to be received in email by one staff member, assigned to another to review, and a third to categorize. The process should

eliminate the potential for extra, unnecessary communication regarding a document.

- Minimize the amount of communication required to add additional state agencies to the repository.

The State Library process should require minimal setup to receive documents from additional agencies. The system should be rules- based allowing staff to customize what documents are selected and how they are processed with minimal, manual intervention.

The State Library process should allow for the agencies to do most of the work in submitting a document by posting it to their website or emailing it to the library. The system should support a self-service approach.

- High growth in number of documents requires additional staff commitment

If the project is highly successful or statutory changes require the agencies to submit more documents, it may require additional staff commitment. An automated process will not require as many staff hours and additional staff to adjust to a higher level of documents (input).

Technology

- Technical requirements for maintaining electronic storage

The State Library project will accumulate a large amount of data over the course of the project. The technical acumen of supporting a large scale of data may require extensive knowledge of the hardware and software products involved to maintain an efficient system. The State Library may not have the level of technical staff available.

The State Library may not be able to get full application support from any hosting facility to adequately cover risks involved. The hosting facility selected may not be trained with the specific software the State Library chooses. The State Library may be required to train other external groups for maintenance of the software or may not be able to find adequate resources.

- Difficulty in maintaining full text of documents

The software package will be required to import and index a very large volume of text. All solutions require that the original data be stored and an indexed version of it. This requires indexes that range from twice (2x) to four (4x) times the size of the physical documents.

- Revisions of documents

The State Library will want to store every version of a document when the project is fully implemented. This may result in much larger data volumes than expected.

- Keyword Indexes change over time with cultural changes

The State Library will use taxonomy standards but over the expected life of the project (50+ years), the meaning of these taxonomies will change and likely result in new taxonomies. The State Library may choose to store multiple taxonomy structures and duplicate sets of documents. This may result in much larger data volumes than expected.

- Redundancy of storage

The State Library will require adequate redundancy to prevent system failures. It's essential for statutory reasons that no document be erased or unrecoverable. The storage costs will be higher to address this issue and the expected electronic storage will be twice to four times the size of the physical documents.

Constraints

- The State Library must follow the state statutes: Massachusetts General Laws (MGL) chapter 6 Section 39 B: the State Library is a depository for every state document. Section 39 applies to every document regardless of its format or manner of duplication.
- The project is designed to support 5 to 10 concurrent users viewing and searching the full text indexes.

Product Requirements

This section contains the specifications and requirements for the hardware, software, and staffing.

Applicable Standards

Security

System level security should follow standard state applicable guidelines for any server products, which require a login including Microsoft Exchange.

Disaster Recovery

The disaster recovery plan will be designed as part of the server implementation plan. It should include regular backups.

Systems Requirements

Hardware Requirements

Optimal Architecture

The optimal architecture for this project is to allow the project to meet new demands simply by adding low cost servers. This can best be met by having a hardware switch which load balances website traffic between multiple web servers. The web servers require local storage to hold the indexes for the application. The web servers directly connect to a storage array which holds all documents in the repository.

Data Storage

The project has two approaches for data storage:

- Network Attached Storage (NAS): The NAS utilizes a lower cost server to keep the required disks. With low cost high capacity drives this is an inexpensive way to create the central document repository. This approach with 250 IDE drives can accommodate roughly 2.5 Terabytes of raw storage. With an appropriate level of redundancy it could offer just over a Terabyte of information. A NAS can be located at the State Library or at the data center.

- Storage Area Network (SAN): The SAN utilizes higher performance equipment including Fibre Channel and SCSI disk drives. This level of

performance would be required once the State Library crosses a threshold of concurrent users. Without further deployment information, this can only be estimated at 50 concurrent users. The application will currently require 5-10 concurrent users and is not a requirement.

Data Center Storage

Long Term and Short Term

The State Library should consider getting the system running in as short a time as possible minimize the risk of documents never being captured. The data storage should be brought online as quickly as possible. In the long term, the State Library must consider the most economical means of storage. Financially, it will be to the State Library's advantage to own the data storage units to minimize ongoing costs.

Ongoing costs are high as needs increase

As the State Library receives additional documents from agencies or other factors increase the size of the repository the operational costs of the system will rise. The State Library will want to carefully monitor its use of disk storage and make operational decisions to buy additional disk storage on a defined interval period.

Additional hardware is required with certain software solutions

With some of the software mentioned in this study it requires additional software. Oracle database is required for Blue Angel, which will require a DBA. It's estimated the Oracle database will be between 600 GB and 1.2 TB, which requires an experienced DBA to keep the system running and properly maintained.

5 year capacity plan allows for up to 600 gigabyte of storage representing approximately 600,000 documents

Staff Requirements

Goal to limit additional work required of existing staff

The software will require staff to become familiar with the categorization methods of the software to properly index documents. Beyond this requirement, the software should allow most functions to be performed without programming knowledge.

Software will require application administrator

All software packages evaluated in this study are sophisticated and complex. They all require an administrator capable of administering Microsoft Windows Server and familiarity with startup/shutdown and troubleshooting issues.

Software will require staff to design/implement indexing or categorization strategy

The software will require staff to undertake and understand the rules based categorization. It's likely this will involve staff training to ensure they have an adequate understanding of the technology.

Location of equipment (factors)

Electronic Storage using Network Attached Storage (NAS) or Storage Area Network (SAN) requires experience with RAID and backup

The storage repository will require knowledge of disk storage servers such as NAS and SAN. It will require knowledge of Redundant Array of Inexpensive Devices (RAID) and backup methods for Windows Server.

Access to Exchange Servers

The software package will likely require access to the State Library's Exchange Server. A software component will need to have access to the Domain and Server in an appropriate fashion to retrieve documents.

Access to Agency Web Sites

The agencies' web sites will need to be accessed through the internet. There are no limitations that would prevent this.

Performance Requirements

The software package will be required to have sub-second response times for full-text search and web browsing.

- The initial project will allow Internal and Public access at the State Library's discretion.
- The project will assume that between 5 to 10 concurrent users